CUNY SPS DATA 621 - CTG5 - HW1

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

February 27, 2019

Contents

1 DATA EXPLORATION

Professionals and gamblers alike are always seeking to optimize their chances of winning, whether it be sports, games, or their bets on them. Major League Baseball is a multibillion dollar industry where individual teams, players, and those who profit off of their success stand to benefit most from such optimization.

Data from 1871 to 2006 was collected in order to infer how many wins could be expected from the 162 games in a baseball team's season. Each observation represents a season for an unnamed team, and we have a total of 2,276 observations. For each team the target variable, `TARGET_WINS`, represents the number of wins in a given year and has a maximum value of 162 possible wins. In addition to that 15 continuous integer predictor variables were collected (not including the index) representing each team's: base hits, doubles, triples, homeruns, walks, and strikeouts by batters, batters hit by pitches, bases stolen by batters and the number of times they were caught stealing, the number of errors, double plays, walks, hits, and homeruns allowed, and strikeouts by pitchers. The testing data contains the same 15 predictor variables and no target variable so it will be impossible to check the accuracy of our predictions from the testing data.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT ON WINS
TARGET_WINS	Number of wins	outcome variable
BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact
BATTING_2B	Doubles by batters (2B)	Positive Impact
BATTING_3B	Triples by batters (3B)	Positive Impact
BATTING_HR	Homeruns by batters (4B)	Positive Impact
BATTING_BB	Walks by batters	Positive Impact
BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact
BATTING_SO	Strikeouts by batters	Negative Impact
BASERUN_SB	Stolen bases	Positive Impact
BASERUN_CS	Caught stealing	Negative Impact
FIELDING_E	Errors	Negative Impact
FIELDING_DP	Double Plays	Positive Impact
PITCHING_BB	Walks allowed	Negative Impact
PITCHING_H	Hits allowed	Negative Impact
PITCHING_HR	Homeruns allowed	Negative Impact
PITCHING_SO	Strikeouts by pitchers	Positive Impact

1.1 Summary Statistics

	n	min	mean	median	max	sd
TARGET_WINS	2276	0	80.79086	82.0	146	15.75215
BATTING_H	2276	891	1469.26977	1454.0	2554	144.59120
BATTING_2B	2276	69	241.24692	238.0	458	46.80141
BATTING_3B	2276	0	55.25000	47.0	223	27.93856
BATTING_HR	2276	0	99.61204	102.0	264	60.54687
BATTING_BB	2276	0	501.55888	512.0	878	122.67086
BATTING_SO	2174	0	735.60534	750.0	1399	248.52642
BASERUN_SB	2145	0	124.76177	101.0	697	87.79117
BASERUN_CS	1504	0	52.80386	49.0	201	22.95634
BATTING_HBP	191	29	59.35602	58.0	95	12.96712
PITCHING_H	2276	1137	1779.21046	1518.0	30132	1406.84293
PITCHING_HR	2276	0	105.69859	107.0	343	61.29875
PITCHING_BB	2276	0	553.00791	536.5	3645	166.35736
PITCHING_SO	2174	0	817.73045	813.5	19278	553.08503
FIELDING_E	2276	65	246.48067	159.0	1898	227.77097
FIELDING_DP	1990	52	146.38794	149.0	228	26.22639

Looking at the above, it can be easily noted that there are outliers present in more than variable, with `PITCHING_H` being the worst offender. Even at three times the standard deviation, its maximum value lays far outiside of the 68-95-99.7 rule. `FIELDING_E`, on the other hand, has the curious case of having a large difference between its mean and median, indicating there is skew present in this variable as well before any charts are actively looked at. Skewed variables cause bias in linear models and need treatment before being used.

1.2 Shape of Predictor Distributions

The distribution of most of the variables seems normal although `BASERUN_SB`, `BASERUN_CS`, and `BATTING_3B` have a slight to moderate right skew, `FIELDING_E`, `PITCHING_BB`, `PITCHING_H`, and `PITCHING_SO` have an extreme right skew, and `BATTING_HR`, `BATTING_SO`, and `PITCHING_HR` are bimodal. As a result some data transformation will most likely be necessary to improve the accuracy of our model. The standard deviation of the various variables also hints at the intense skewing of some of the variables.

1.3 Outliers

There are also a large number of outliers that need to be accounted for, most prevalently in `FIELDING_E` and `BATTING_H` based off of the boxplots above. One such extreme outlier removed implied that there were, on average per game in a single season, 186 hits allowed by pitchers. This is an unrealistic figure, even for those for whom baseball is outside of their realm of understanding.

1.4 Missing Values

Of all the observations gathered across these fifteen variables, there are 3,478 missing values out of 36,416 total data points, which represents 10.187% of the data. Batters hit by pitches was missing the most, with 2,085 instances of missing information, which represents 91.61% of that variable missing. Additionally `Pitching_SO` and `Batting_SO` are missing exact same proportion 4.48% and are missing in the same observations. This data may not be missing at random and so there may be cause for removing it.

1.5 Linearity

Each variable was tested against the target variable in order to determine at a glance which had the most potential linearity before the dataset was modified.

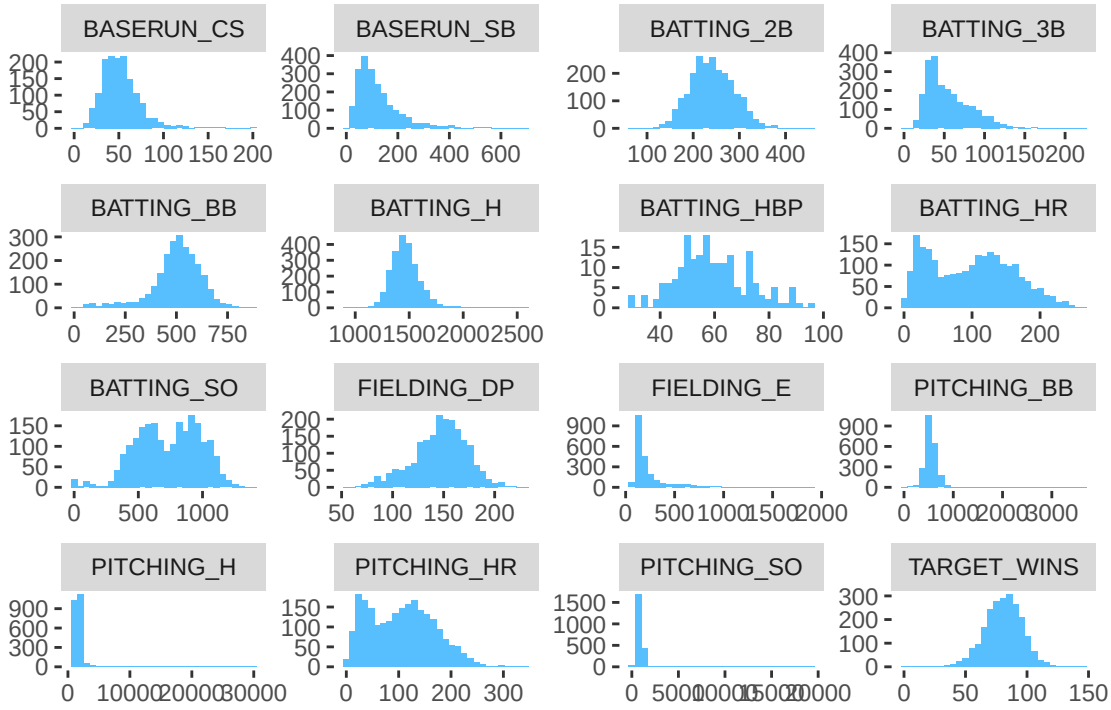


Figure 1: Data Distributions

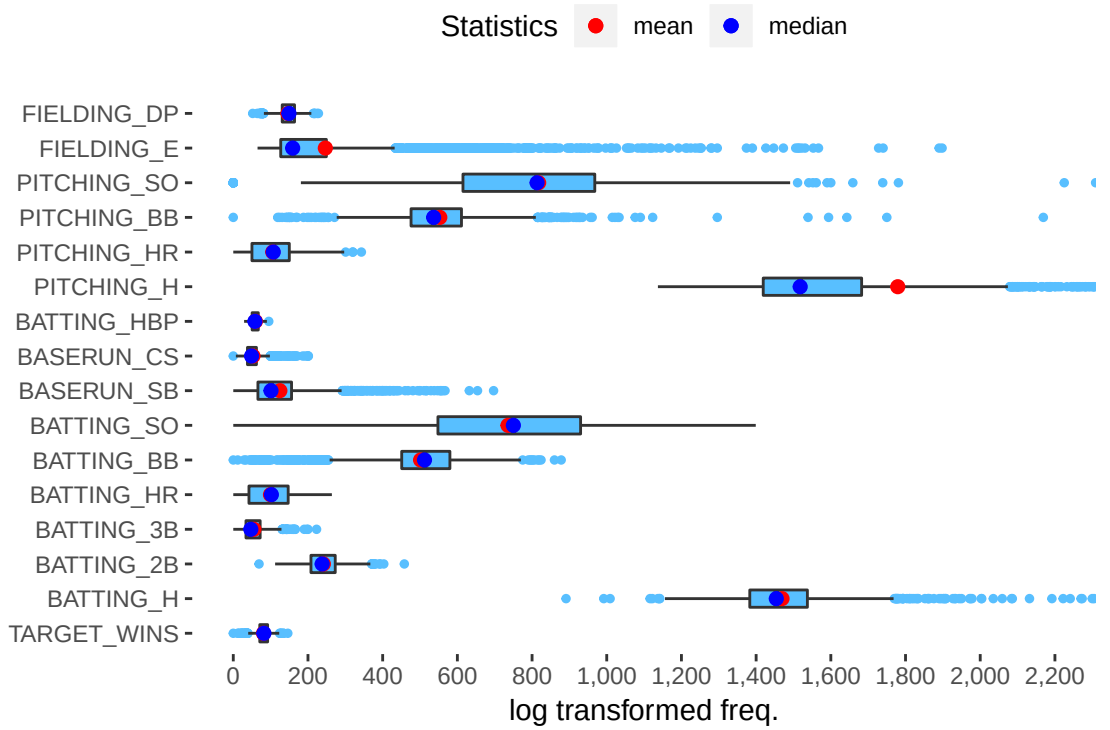


Figure 2: Boxplots highlighting many outliers in the data.

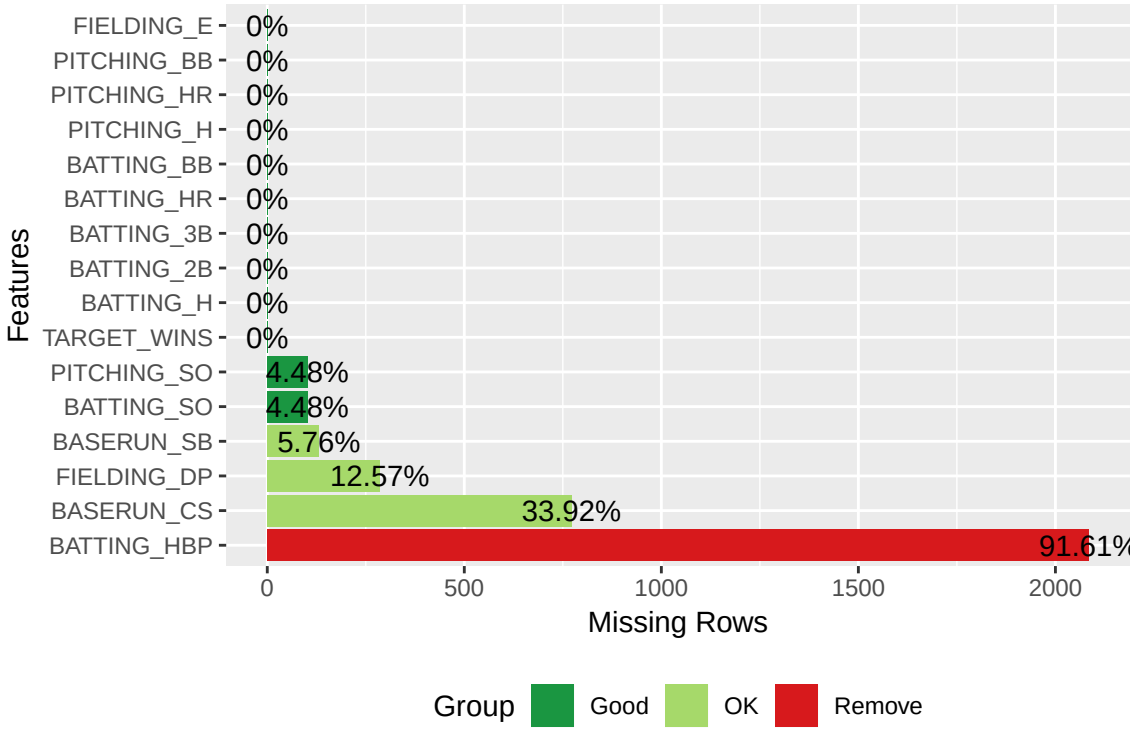


Figure 3: Missing values

As can be observed, the most influential variables are the ones previously discussed to have severe outliers and skew, and their linear relationship is negative - the higher the variable, the lower the target wins. On the other hand, `BATTING_H`, `BATTING_BB` and `BATTING_2B` showed the most promise.

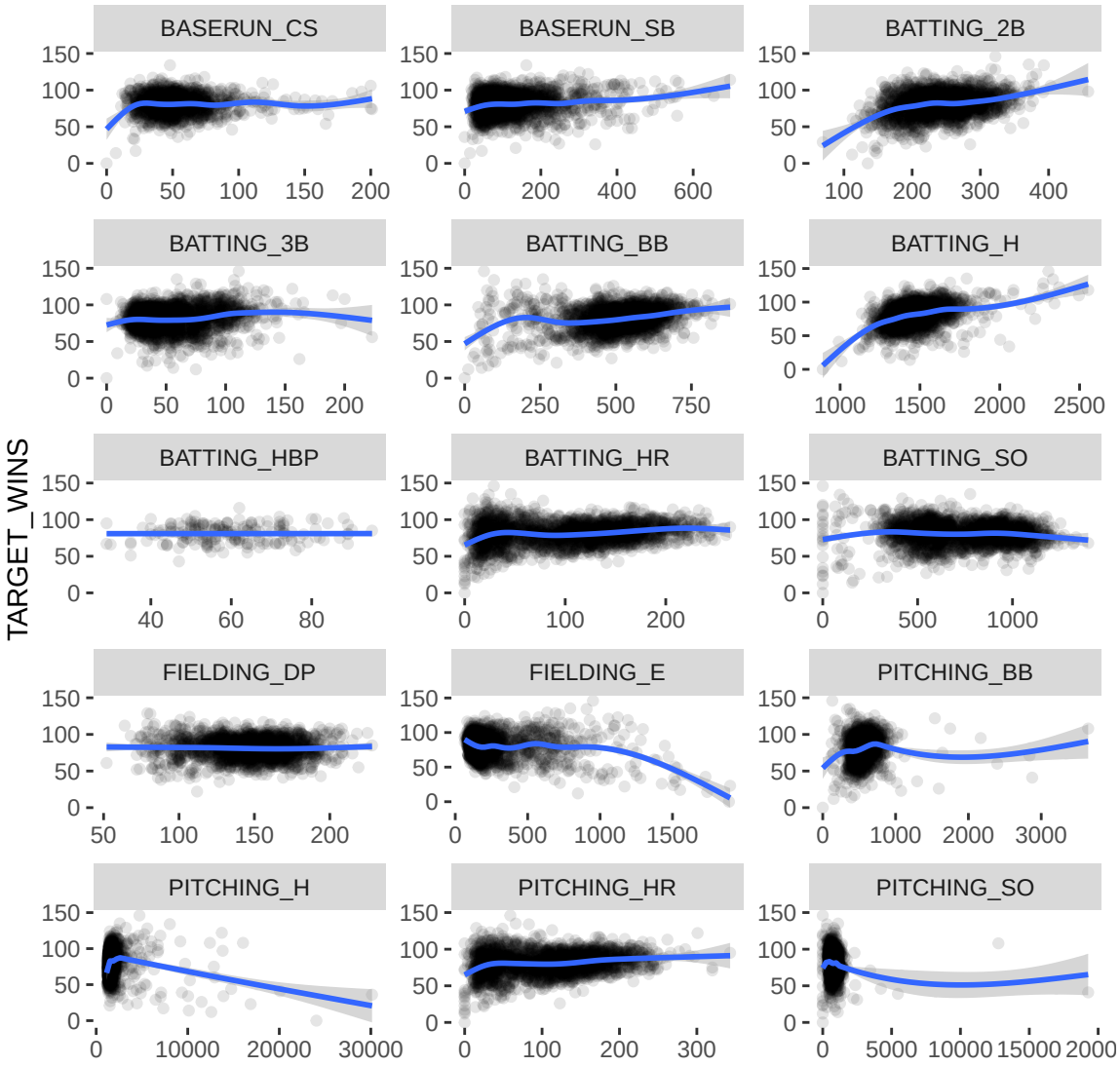


Figure 4: Linear relationships between each predictors and the target

2 DATA PREPARATION

2.1 Missing Values

As previously mentioned, just north of 10% of the data was missing values. Missing values can lead to errors in a model, bias, and worse if left unaccounted for. Attempting to "fix" this by imputing values or guessing why the values are missing in the first place - such as concluding that the missing values are meant to be zeroes - are just as likely to help with creating a model as it is to help with creating a disaster.

One of the R packages utilized, DataExplorer, which was used for the chart above, recommends removing null or missing values above a certain threshold as indicated in the graph.

Fixing missing values with imputation may help, but can also have a negative impact on the model if the assumed values do not correspond to the actual missing values. When it is just a few observations missing, modifications can be made, however, 91.61% is too large a proportion and would almost definitely distort the model, so we decided it was better to remove the `BATTING_HBP` column altogether. Deleting all cases with missing values, in this instance, would have shrunk the size of the dataset down to less than a tenth of its original size. If we simply delete all cases with missing values from the analysis, we will cause no bias, but we would most certainly lose a lot of important information.

Data that is Missing Completely at Random (MCAR), meaning the probability that a value is missing is the same for all cases can be imputed. Although there is some concern about whetherr or not `Pitching_SO` and `Batting_SO` are MCAR, we chose to leave all the remaining variables except `BATTING_HBP` and determine whether or not to remove themn during the modelling process.

2.1.1 NA Imputation

To deal with the remaining missing values, the bag imputation method was used via the caret package. A set of dummy variables were created and were used to predict the various values in the dataset. This dummy-set was then pre-processed and used against itself to predict the missing values.

	n	min	mean	median	max	sd
TARGET_WINS	2276	0	80.79086	82.0	146	15.75215
BATTING_H	2276	891	1469.26977	1454.0	2554	144.59120
BATTING_2B	2276	69	241.24692	238.0	458	46.80141
BATTING_3B	2276	0	55.25000	47.0	223	27.93856
BATTING_HR	2276	0	99.61204	102.0	264	60.54687
BATTING_BB	2276	0	501.55888	512.0	878	122.67086
BATTING_SO	2174	0	735.60534	750.0	1399	248.52642
BASERUN_SB	2145	0	124.76177	101.0	697	87.79117
BASERUN_CS	1504	0	52.80386	49.0	201	22.95634
PITCHING_H	2276	1137	1779.21046	1518.0	30132	1406.84293
PITCHING_HR	2276	0	105.69859	107.0	343	61.29875
PITCHING_BB	2276	0	553.00791	536.5	3645	166.35736
PITCHING_SO	2174	0	817.73045	813.5	19278	553.08503
FIELDING_E	2276	65	246.48067	159.0	1898	227.77097
FIELDING_DP	1990	52	146.38794	149.0	228	26.22639

2.2 Remove Outliers

Outlier treament was done by placing a threshold of five times the standard deviation up from the mean and removing all observations that fell north of this boundary.

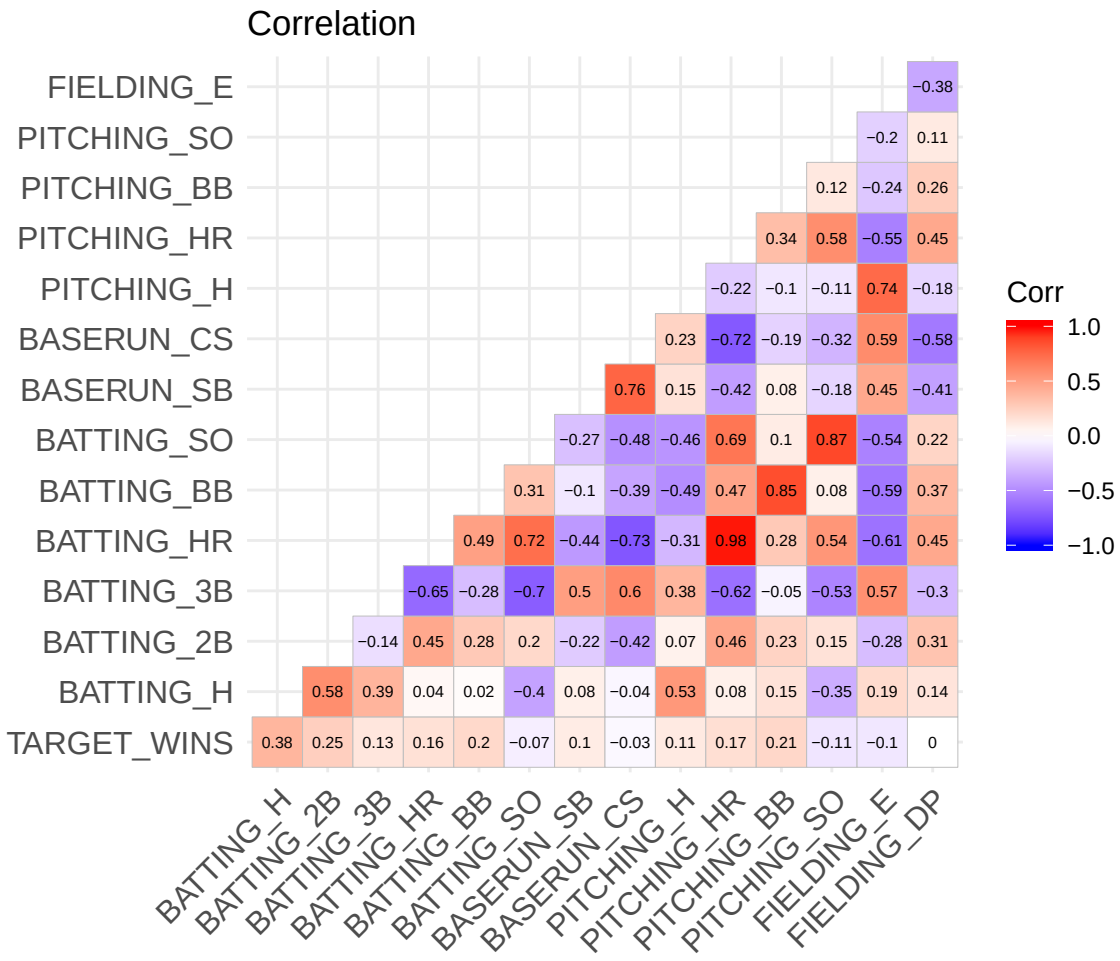
```
## TARGET_WINS   BATTING_H  BATTING_2B  BATTING_3B  BATTING_HR  BATTING_BB
```

```
##           0           0           0           0           0           0
##  BATTING_SO  BASERUN_SB  BASERUN_CS  PITCHING_H PITCHING_HR PITCHING_BB
##         102         131         772           0           0           0
## PITCHING_SO  FIELDING_E FIELDING_DP
##         102           0         286
```

2.3 Correlation

The theoretical effect of strikeouts by batters, batters caught stealing, errors, walks, hits, and homeruns allowed were believed to have a negative impact on the number of wins of an individual team in a given year. A closer look at the correlation plot between the variables painted a different picture.

When compared to what was hypothesized, there was actually a positive impact for the number of wins for a team in a given year by walks, hits, and homeruns allowed; at the same time, variables previously thought to have a positive correlation - strikeouts by pitchers and double plays - had a negative correlation for the number of wins. The three variables with the greatest correlation to the number of wins were the hits allowed, the walks by batters, and the walks allowed. Of these, the hits allowed had a relatively low correlation with the walks by batters and the walks allowed, whereas the walks allowed and the walks by batters had a direct positive correlation with one another.



2.4 Feature Engineering

Jeremy: Adjusted this to reflect offense (batting) minus defense (pitching). These arithmetically transformed offense / defense variables are linearly related with BATTING and PITCHING variables, so we can include one or the other in a model, but not both. Replacing original variables with these transforms did not improve R^2 in a base case.

FOR THE OTHER HALF OF THE GROUP:

z_train <- sapply(imputed_train, scale) log_train <- log(imputed_train) # weird results z_log_train <- sapply(log_train, scale) # weirder results

imputed_train is most likely the variable you want to use.

---

3 BUILD MODELS

3.1 Instructions:

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

3.2 MODEL 1

Multiple regression can be created as a purely statistical model, through the use of significance tests, or it can be interpreted in a more practical, non-statistical manner. This approach is based on the subject-area expertise.

We've created the following categories from the most important to the least important variables according to the subject-area expert.

Very Important: BATTING_H, BATTING_HR, BATTING_SO ,FIELDING_E, PITCHING_SO

Fairly Important: BASERUN_SB, PITCHING_HR, BATTING_BB

Important: BATTING_2B, BATTING_3B, FIELDING_DP, PITCHING_H

Slightly Important: PITCHING_BB, BASERUN_CS

Not at all important: BATTING_HBP

'Batters hit by pitch' and 'Caught Stealing' have been eliminated as least important variables according to the expert.

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO +
##     FIELDING_E + PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB +
##     BATTING_2B + BATTING_3B + FIELDING_DP + PITCHING_BB + PITCHING_H,
##     data = imputed_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.602  -8.272   0.065   7.985  69.047
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.592272   5.663910   6.461 1.28e-10 ***
## BATTING_H    0.016719   0.005010   3.337 0.000861 ***
## BATTING_HR   0.119188   0.050659   2.353 0.018723 *
## BATTING_SO  -0.031132   0.006598  -4.718 2.53e-06 ***
## FIELDING_E  -0.038259   0.003433 -11.144  < 2e-16 ***
```

```
## PITCHING_SO  0.019966   0.005420   3.684 0.000235 ***
## BASERUN_SB   0.041446   0.004934   8.399  < 2e-16 ***
## PITCHING_HR -0.034985   0.046876  -0.746 0.455549
## BATTING_BB   0.078335   0.016217   4.830 1.46e-06 ***
## BATTING_2B  -0.011063   0.009390  -1.178 0.238836
## BATTING_3B   0.126375   0.018155   6.961 4.43e-12 ***
## FIELDING_DP -0.091226   0.013111  -6.958 4.53e-12 ***
## PITCHING_BB -0.055706   0.014380  -3.874 0.000110 ***
## PITCHING_H   0.013513   0.001539   8.779  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.63 on 2216 degrees of freedom
## Multiple R-squared:  0.2835, Adjusted R-squared:  0.2793
## F-statistic: 67.46 on 13 and 2216 DF,  p-value: < 2.2e-16
```

We got 0.2793 on Adjusted R-squared after we removed these two variables. Once we tried to remove other not very important variables according to subject-area expert, we got an even lower R-squared.

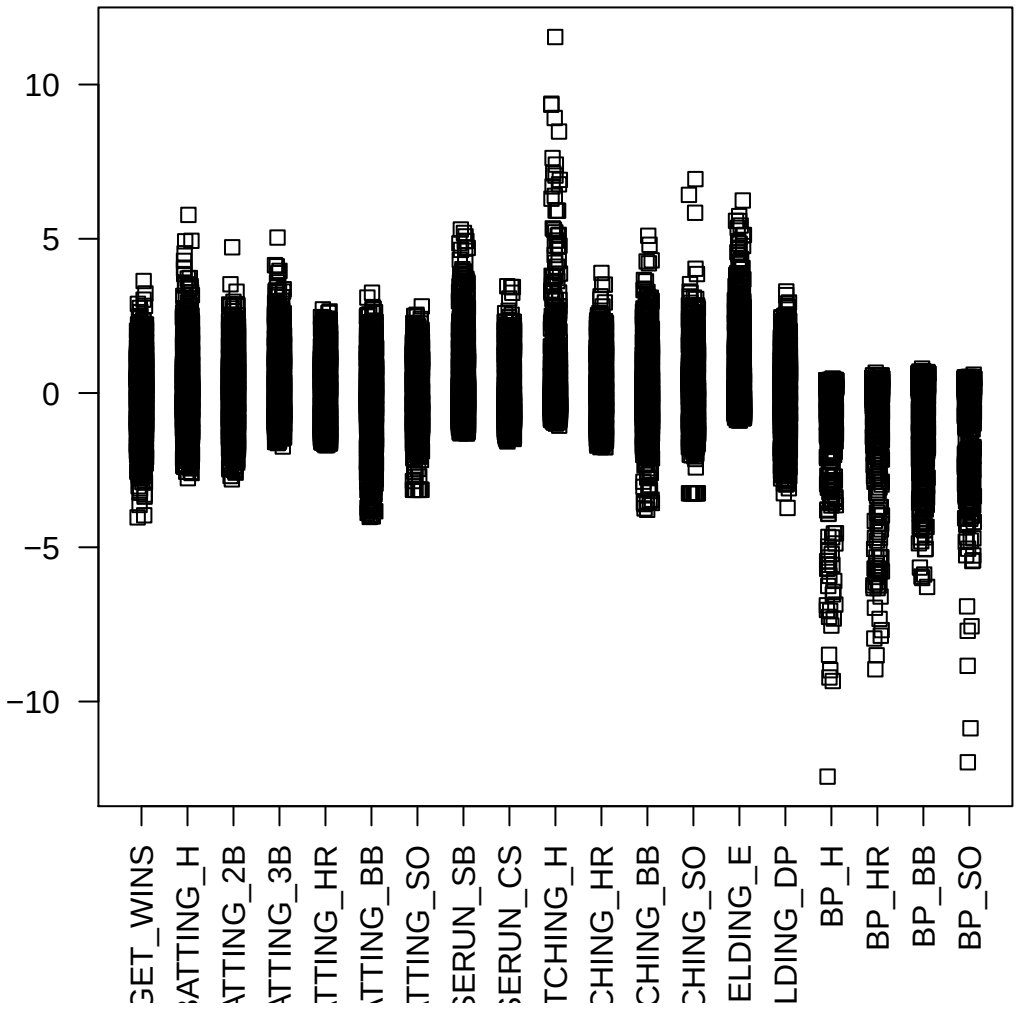
The next step we performed was backward elimunation, which was more effective compared to forward selection. BATTING_H and BATTING_2B have been removed based on the Backward Selection results.

```
## Start:  AIC=11325.02
## TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO + FIELDING_E +
##     PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB + BATTING_2B +
##     BATTING_3B + FIELDING_DP + PITCHING_BB + PITCHING_H
##
##               Df Sum of Sq    RSS   AIC
## - PITCHING_HR  1      88.9 353605 11324
## - BATTING_2B   1     221.5 353737 11324
## <none>                     353516 11325
## - BATTING_HR   1     883.1 354399 11329
## - BATTING_H    1    1776.5 355292 11334
## - PITCHING_SO  1    2165.1 355681 11337
## - PITCHING_BB  1    2393.9 355910 11338
## - BATTING_SO   1    3551.6 357068 11345
## - BATTING_BB   1    3722.3 357238 11346
## - FIELDING_DP  1    7723.5 361239 11371
## - BATTING_3B   1    7730.1 361246 11371
## - BASERUN_SB   1   11255.0 364771 11393
## - PITCHING_H   1   12294.0 365810 11399
## - FIELDING_E   1   19811.3 373327 11445
##
## Step:  AIC=11323.58
## TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO + FIELDING_E +
##     PITCHING_SO + BASERUN_SB + BATTING_BB + BATTING_2B + BATTING_3B +
##     FIELDING_DP + PITCHING_BB + PITCHING_H
##
##               Df Sum of Sq    RSS   AIC
## - BATTING_2B   1     218.7 353824 11323
## <none>                     353605 11324
## - BATTING_H    1    1790.1 355395 11333
## - PITCHING_SO  1    2131.4 355736 11335
## - BATTING_SO   1    3491.5 357096 11344
## - PITCHING_BB  1    4558.7 358164 11350
```

```
## - BATTING_BB   1    6449.9 360055 11362
## - BATTING_3B   1    7663.0 361268 11369
## - FIELDING_DP  1    7822.8 361428 11370
## - BATTING_HR   1   11191.6 364796 11391
## - BASERUN_SB   1   11302.9 364908 11392
## - PITCHING_H   1   12384.3 365989 11398
## - FIELDING_E   1   20640.3 374245 11448
##
## Step:  AIC=11322.96
## TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO + FIELDING_E +
##     PITCHING_SO + BASERUN_SB + BATTING_BB + BATTING_3B + FIELDING_DP +
##     PITCHING_BB + PITCHING_H
##
##               Df Sum of Sq    RSS   AIC
## <none>                      353824 11323
## - BATTING_H    1    1777.1 355601 11332
## - PITCHING_SO  1    1990.8 355814 11334
## - BATTING_SO   1    3479.4 357303 11343
## - PITCHING_BB  1    4495.2 358319 11349
## - BATTING_BB   1    6351.2 360175 11361
## - FIELDING_DP  1    7773.8 361597 11369
## - BATTING_3B   1    8094.2 361918 11371
## - BATTING_HR   1   11452.8 365276 11392
## - BASERUN_SB   1   11828.7 365652 11394
## - PITCHING_H   1   12901.1 366725 11401
## - FIELDING_E   1   20437.9 374261 11446
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO +
##     FIELDING_E + PITCHING_SO + BASERUN_SB + BATTING_BB + BATTING_3B +
##     FIELDING_DP + PITCHING_BB + PITCHING_H, data = imputed_train)
##
## Coefficients:
## (Intercept)    BATTING_H   BATTING_HR   BATTING_SO   FIELDING_E
##    38.93008      0.01316      0.08286     -0.03074     -0.03727
## PITCHING_SO   BASERUN_SB   BATTING_BB   BATTING_3B  FIELDING_DP
##     0.01896      0.04219      0.08441      0.12819     -0.09140
## PITCHING_BB   PITCHING_H
##    -0.06157      0.01375
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_HR + BATTING_SO + FIELDING_E +
##     PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB + BATTING_3B +
##     FIELDING_DP + PITCHING_BB + PITCHING_H, data = imputed_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.349  -8.319   0.095   8.044  73.821
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.967361   3.815685  13.357  < 2e-16 ***
## BATTING_HR   0.137603   0.050478   2.726 0.006461 **
```

```
## BATTING_SO  -0.033187   0.006582  -5.042 4.98e-07 ***
## FIELDING_E  -0.041649   0.003245 -12.835  < 2e-16 ***
## PITCHING_SO  0.018692   0.005384   3.472 0.000527 ***
## BASERUN_SB   0.045791   0.004785   9.569  < 2e-16 ***
## PITCHING_HR -0.038873   0.046964  -0.828 0.407915
## BATTING_BB   0.085893   0.016055   5.350 9.70e-08 ***
## BATTING_3B   0.156437   0.016039   9.754  < 2e-16 ***
## FIELDING_DP -0.083573   0.012950  -6.454 1.34e-10 ***
## PITCHING_BB -0.061960   0.014270  -4.342 1.48e-05 ***
## PITCHING_H   0.016931   0.001185  14.286  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 2218 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.276
## F-statistic: 78.23 on 11 and 2218 DF,  p-value: < 2.2e-16
```

Our R-squared was still low (0.276), so we decided to look at the outliers, which can affect our model. Pitching_h had the high number of outliers which indicated a need for data transformation. We decided to use log transformation for this variable.




```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_HR + BATTING_SO + FIELDING_E +
##     PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB + BATTING_3B +
##     FIELDING_DP + PITCHING_BB + log(PITCHING_H), data = imputed_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.238  -8.139   0.185   8.020  67.777
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.037e+02  2.426e+01 -12.522  < 2e-16 ***
## BATTING_HR       1.134e-01  5.003e-02   2.268   0.0234 *
## BATTING_SO      -1.544e-02  6.668e-03  -2.315   0.0207 *
## FIELDING_E      -4.202e-02  3.159e-03 -13.299  < 2e-16 ***
## PITCHING_SO      7.688e-03  5.336e-03   1.441   0.1498
```

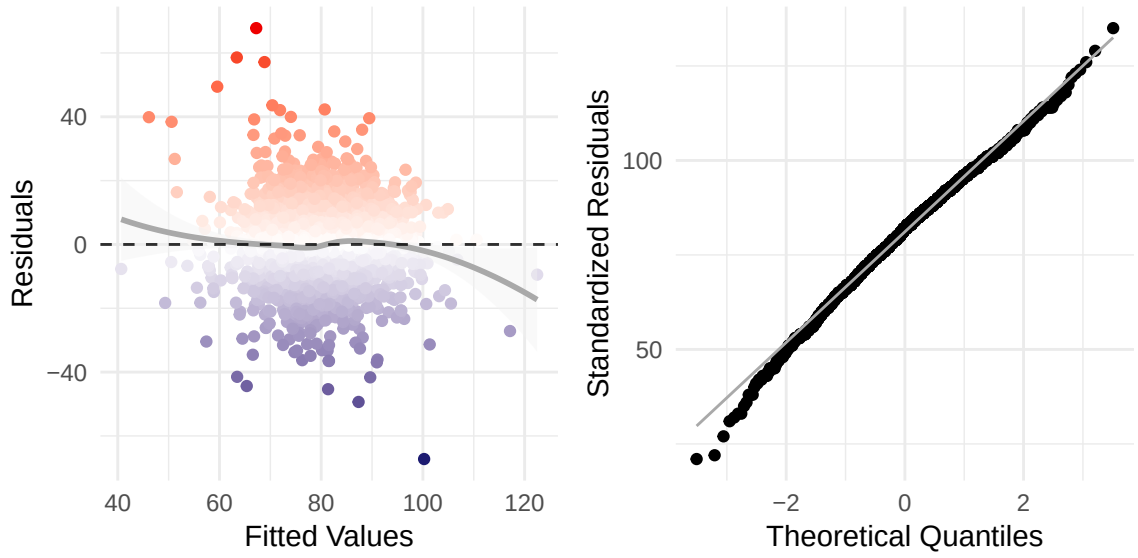


Figure 5: Model 1: Residual Plot and Q-Q Plot

```
## BASERUN_SB       4.138e-02  4.680e-03   8.842  < 2e-16 ***
## PITCHING_HR     -4.352e-02  4.640e-02  -0.938   0.3484
## BATTING_BB       1.011e-01  1.600e-02   6.319 3.16e-10 ***
## BATTING_3B       1.171e-01  1.594e-02   7.349 2.79e-13 ***
## FIELDING_DP     -9.430e-02  1.283e-02  -7.351 2.76e-13 ***
## PITCHING_BB     -7.559e-02  1.423e-02  -5.311 1.20e-07 ***
## log(PITCHING_H)  5.222e+01  3.243e+00  16.099  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.52 on 2218 degrees of freedom
## Multiple R-squared:  0.2956, Adjusted R-squared:  0.2921
## F-statistic:  84.6 on 11 and 2218 DF,  p-value: < 2.2e-16
```

After we used the log trasformation our model's Adjusted R-squared increased to 0.2921.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   60.82   76.09   81.12   81.44   86.10  108.26      54
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.
value	0.2955663	0.2920727	12.51838	84.6026	0	12	-8793.869	17613.74	17687.96	347582.5	

Summary of Results for Model 1:

The overall Subject-Area expertise wasn't as effective as a stand alone method of creating multiple regression models. Statistical iterations which were performed contradicted the subject area expert, such as, removing Batting_H from the model. Additionally log tranformation of PITCHING_H made a significant improvment in our model linearity.

3.3 MODEL 2

Our approach for Model 2 was to try to use as many of the tools as possible that are available in R and that we have learned thus far to determine a model based solely on the statistical qualities of the predictor

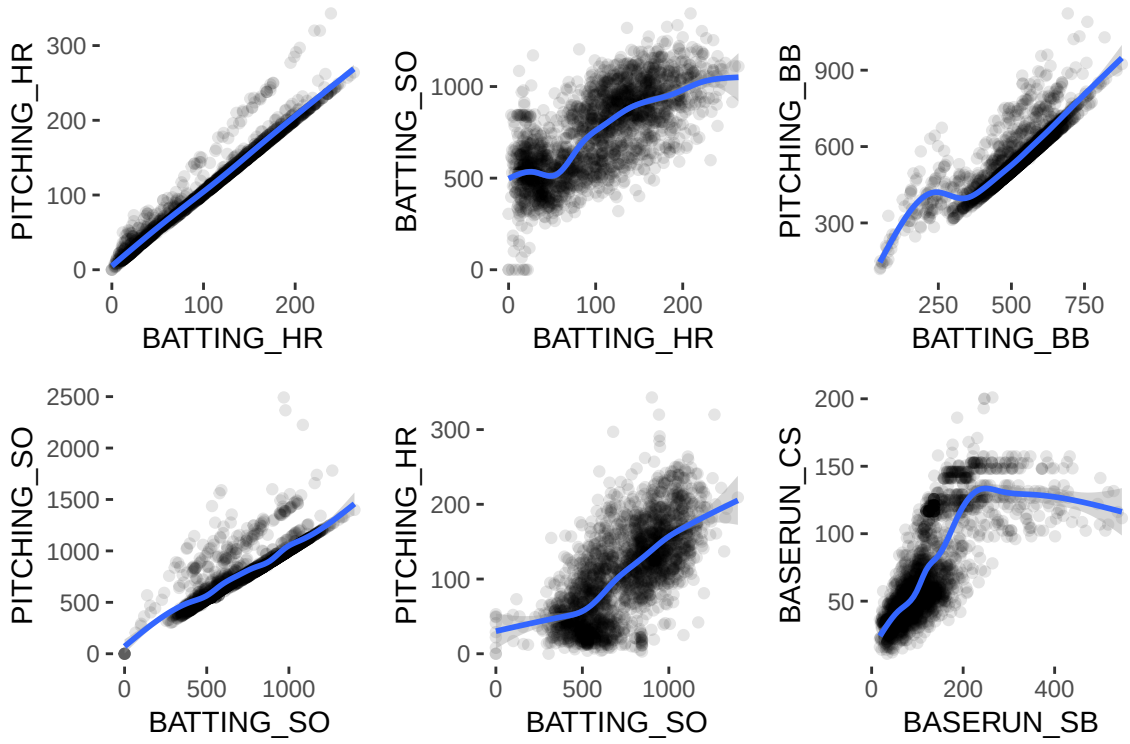


Figure 6: Scatterplots showing possible collinearity problems

variables without any regard to our expert's opinion.

We started by plotting the relationships between variables that had high correlation values to look for potential collinearity problems.

Based on the charts above we decided somewhat arbitrarily to remove the three pitching variables (`PITCHING_HR`, `PITCHING_BB`, and `PITCHING_SO`) rather than the corresponding batting variables (`BATTING_HR`, `BATTING_BB`, and `BATTING_SO`) due to the extremely high correlation between these predictors. We then plotted the reamining variabes to see if they showed a linear relationship with the target variable. Most of the premaining predictors showed a clear linear relationship with the target, however, the extreme skew of `PITCHING_H` and `FIELDING_E` as well as a more moderate skew in `BASERUN_SB` and `BATTING_3B`, can be seen in the plots.

3.3.1 Log Transform Data

We decided to log transform `PITCHING_H`, `FIELDING_E`, `BASERUN_SB` and `BATTING_3B` in order to compensate for the skew. The resulting distributions can be seen in the revised plots below.

3.3.2 Building the Model

Finally we built a model based on the selected variables including the log transformations where appropriate.

```
TARGET_WINS ~ BATTING_H + BATTING_2B + log(BATTING_3B) + BATTING_HR +
    BATTING_BB + BATTING_SO + log(BASERUN_SB) + log(PITCHING_H) +
    log(FIELDING_E) + FIELDING_DP
```

All of the variables had a very low p-value indicating a significant impact on our target, however our R^2 value was low at only 0.2889.

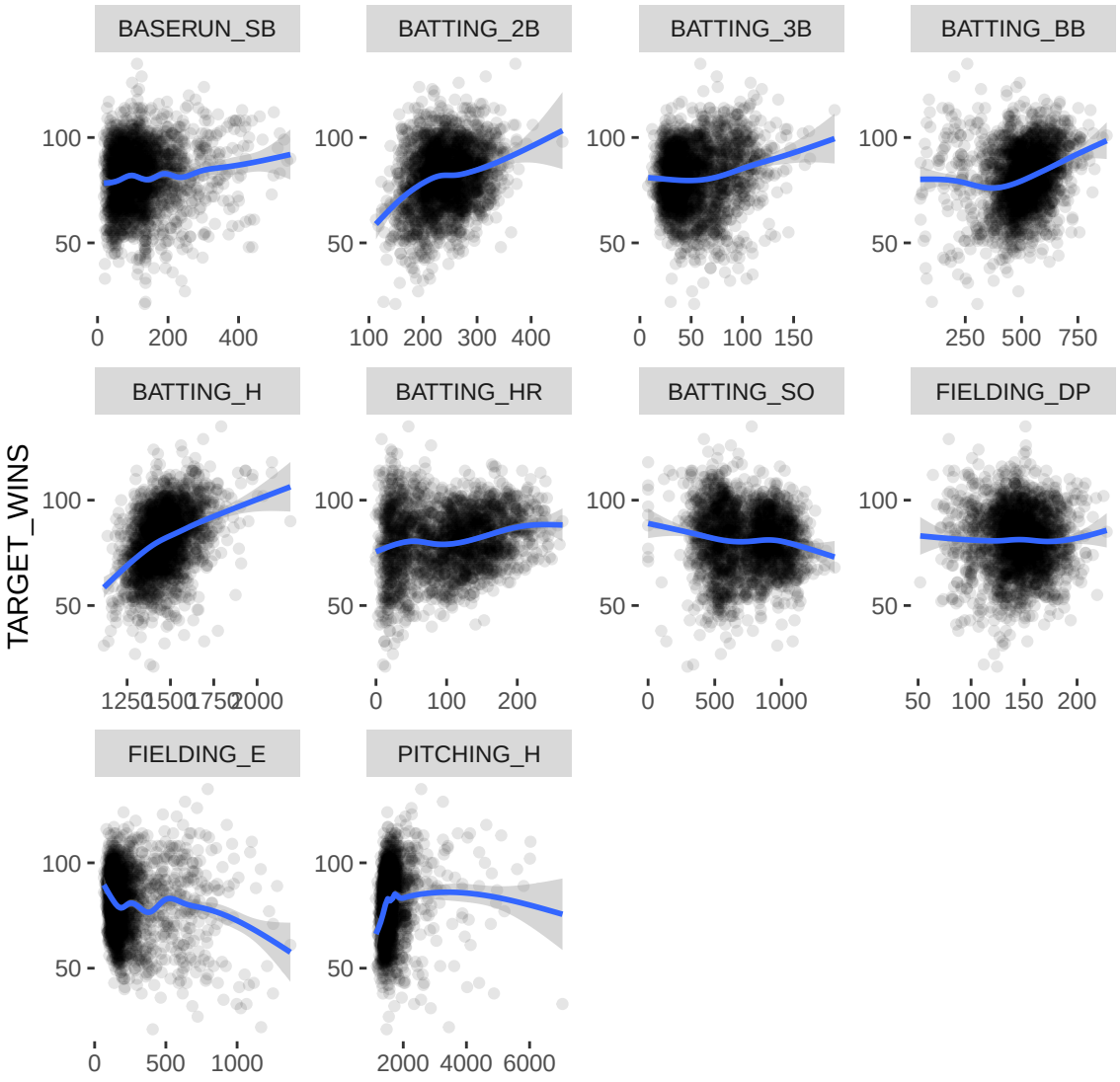


Figure 7: Linear relationship between each predictor and the arget

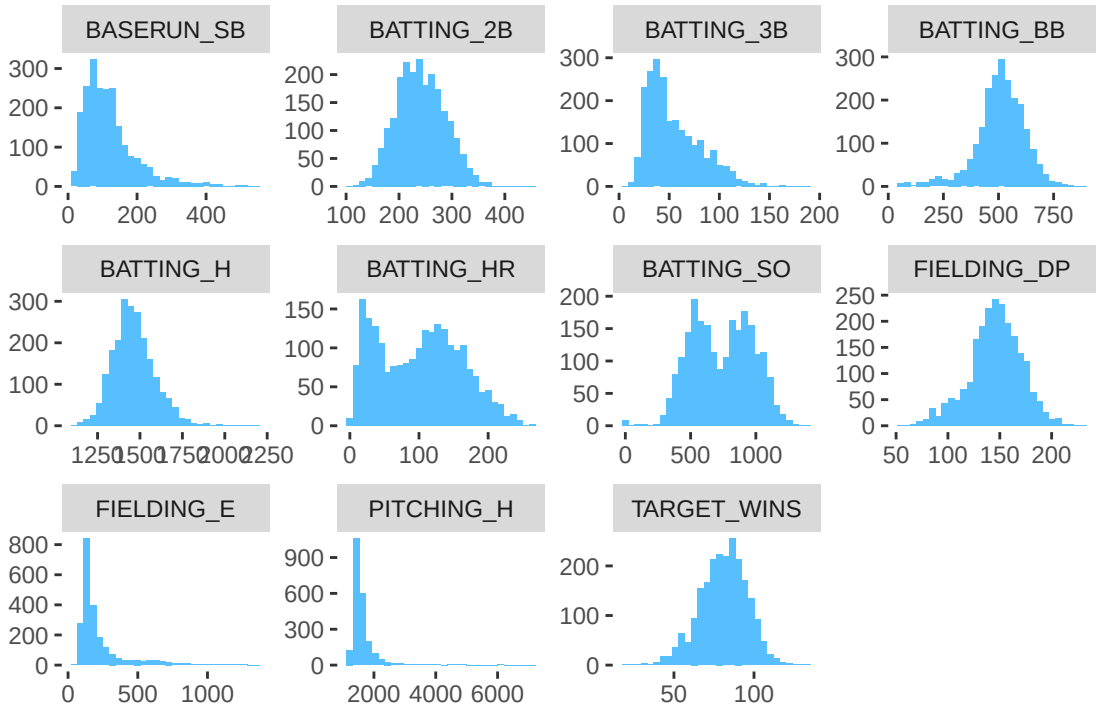


Figure 8: Predictor variable distributions

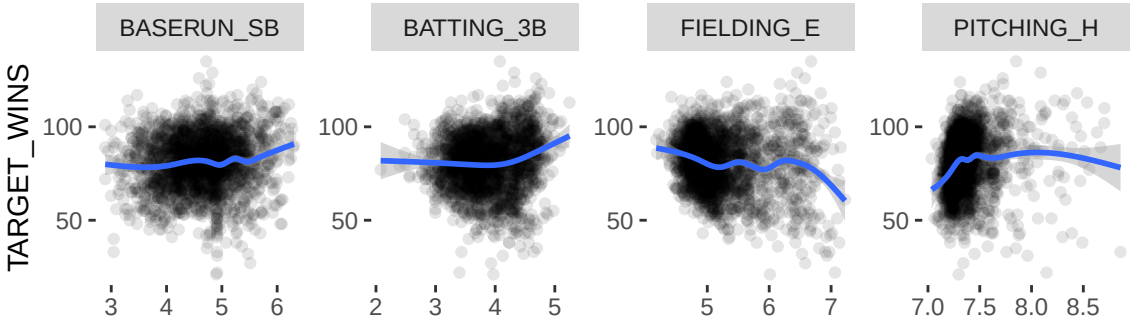


Figure 9: Linear relationship between each log transformed predictor and the Target showing decreased skew

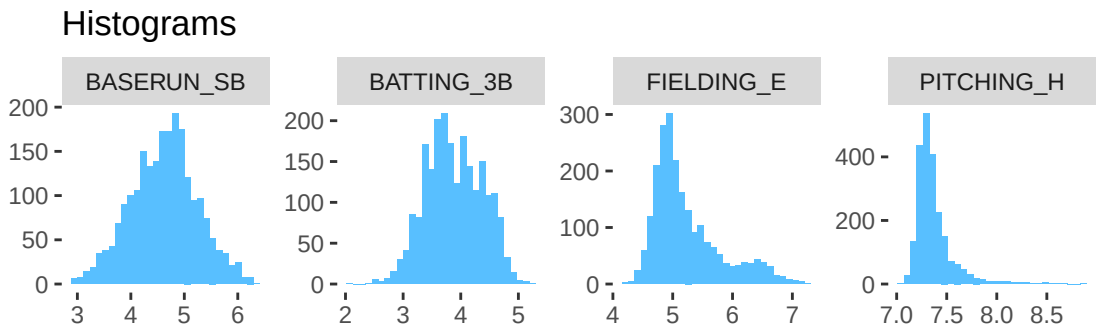


Figure 10: Log transformed distributions showing decreased skew

Table 2: Full Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-62.9724524	15.9139042	-3.957071	0.0000783
BATTING_H	0.0255615	0.0046111	5.543426	0.0000000
BATTING_2B	-0.0297598	0.0092155	-3.229309	0.0012590
log(BATTING_3B)	6.9697812	0.9476551	7.354765	0.0000000
BATTING_HR	0.0703282	0.0101256	6.945579	0.0000000
BATTING_BB	0.0192290	0.0031872	6.033192	0.0000000
BATTING_SO	-0.0112281	0.0024113	-4.656450	0.0000034
log(BASERUN_SB)	4.5460080	0.5617657	8.092356	0.0000000
log(PITCHING_H)	19.1408576	2.6161541	7.316411	0.0000000
log(FIELDING_E)	-13.1027616	1.0618606	-12.339437	0.0000000
FIELDING_DP	-0.1067225	0.0131384	-8.122924	0.0000000

3.3.2.1 R^2 0.2888829

Our residuals look normally distributed and random, and with constant variability, no indication of homoscedasticity. However we thought we may be able to use some other tools in R to refine our model and get a better R^2 value. So next we tried using the leaps package to see if it would recommend removing any of our chosen variables from the model. In the following plot you can see that we could remove `BATTING_H`, `BATTING_2B` without affecting out R^2 much, but it would not improve the model.

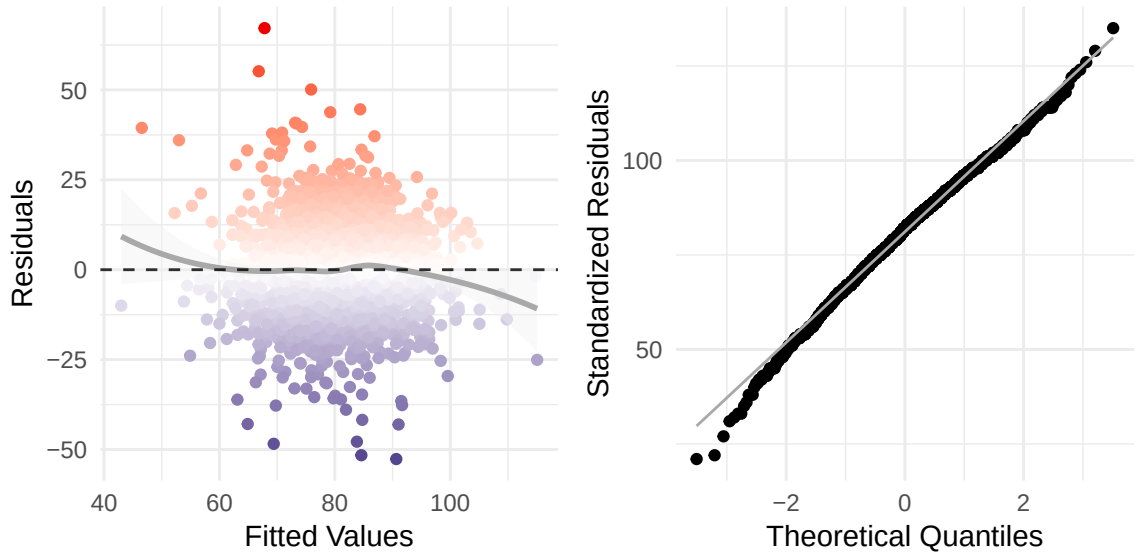


Figure 11: Model 2: Residual Plot and Q-Q Plot

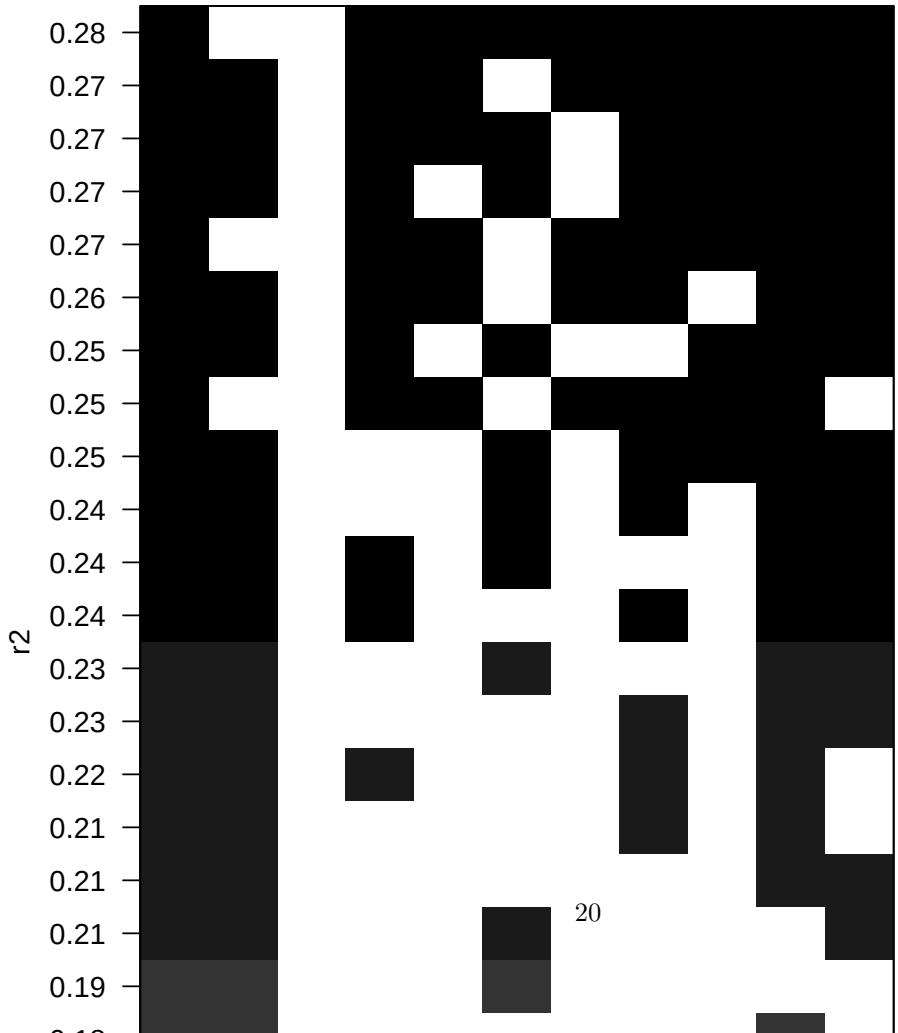


Table 3: Full SCALED Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	80.9717489	0.2690639	300.938712	0.0000000
BATTING_H	2.6935751	0.5977789	4.505972	0.0000070
BATTING_2B	-0.4725422	0.4272176	-1.106092	0.2688064
BATTING_3B	3.0938480	0.4776884	6.476708	0.0000000
BATTING_HR	4.8934200	0.5902167	8.290887	0.0000000
BATTING_BB	1.7865264	0.3813959	4.684178	0.0000030
BATTING_SO	-1.9686032	0.5564927	-3.537519	0.0004122
BASERUN_SB	2.6367694	0.3743523	7.043551	0.0000000
PITCHING_H	4.5352452	0.5648773	8.028726	0.0000000
FIELDING_E	-6.4017884	0.5957772	-10.745273	0.0000000
FIELDING_DP	-2.3837076	0.3295935	-7.232265	0.0000000

Table 4: Subspace Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-47.5598798	14.1082067	-3.371079	0.0007616
I(BATTING_HR + PITCHING_HR)	0.0361009	0.0049624	7.274907	0.0000000
I(BATTING_BB + PITCHING_BB)	0.0077316	0.0014542	5.316742	0.0000001
I(BATTING_SO + PITCHING_SO)	-0.0052127	0.0010862	-4.798948	0.0000017
BATTING_H	0.0273198	0.0048717	5.607838	0.0000000
BATTING_2B	-0.0248999	0.0093466	-2.664069	0.0077761
log(BATTING_3B)	6.4554813	0.9412999	6.858050	0.0000000
log(BASERUN_SB)	3.0279771	0.7904181	3.830855	0.0001312
BASERUN_CS	0.0460357	0.0177169	2.598400	0.0094279
log(PITCHING_H)	17.7088157	2.6101829	6.784511	0.0000000
log(FIELDING_E)	-13.7380684	1.0992719	-12.497426	0.0000000
FIELDING_DP	-0.0968425	0.0134958	-7.175776	0.0000000

```
## NULL
```

Next we tried standardizing the (non-log-transformed) variables to see what impact that might have on our model. As you can see standardzing actually resulted in a significant reduction in our R^2 value from 0.2889 to 0.274.

3.3.2.2 R^2 0.2739688

3.3.3 Test all of the predictors

Next we ran an ANOVA test to compare our model to the null model. With a p-value that is basically zero, clearly our model is statistically significant.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2229	493421.2	NA	NA	NA	NA
2219	350880.3	10	142541	90.14425	0

3.3.4 Testing a subspace

We then tried testing a subspace. Since our initial models using the difference between the corresponding batting and pitching variables did not show promise we tried adding those two variables instead.

3.3.4.1 R^2 0.286633

Table 5: Full Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	204.6302089	103.8266135	1.9708840	0.0488615
BATTING_3B	0.1602499	0.0178747	8.9651836	0.0000000
BATTING_HR	0.0721476	0.0102070	7.0684573	0.0000000
BATTING_BB	0.0624848	0.0199429	3.1331835	0.0017518
PITCHING_BB	-0.0863035	0.0130572	-6.6096358	0.0000000
BATTING_SO	0.0016687	0.0091718	0.1819433	0.8556439
PITCHING_SO	0.0285700	0.0064784	4.4100525	0.0000108
BASERUN_SB	0.0304137	0.0059322	5.1268671	0.0000003
BASERUN_CS	0.0630545	0.0157254	4.0097229	0.0000628
BATTING_H	-0.2185464	0.0544015	-4.0172845	0.0000608
log(PITCHING_H)	-8.4383898	14.3385415	-0.5885110	0.5562494
log(FIELDING_E)	-16.4657197	1.1044063	-14.9091144	0.0000000
FIELDING_DP	-0.0883833	0.0131868	-6.7024185	0.0000000
BATTING_BB:PITCHING_BB	0.0000469	0.0000169	2.7798125	0.0054850
BATTING_SO:PITCHING_SO	-0.0000271	0.0000044	-6.1888757	0.0000000
BATTING_H:log(PITCHING_H)	0.0303947	0.0074285	4.0916590	0.0000444

Once again our model declined in performance rather than improving.

Last, but not least, we used the stepAIC function from the MASS package to see if it came up with different recommendations for what variabels to keep and which to exclude from our model. We started with all variables putting back the ones we had previously taken out due to collinearity issues and let the algorithm choose which to keep.

The final suggested model was:

```
Final Model:
TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR + BATTING_BB +
    BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H + PITCHING_BB +
    PITCHING_SO + FIELDING_E + FIELDING_DP
```

In comparison to our original model we had the following variables added to our model (`BASERUN_CS`, `PITCHING_BB`, and `PITCHING_SO`) and the following variable removed (`BATTING_2B`).

We tried multiple iterations of that model, without any log transformations, with log transformations, with and without the collinear variables, but whever we removed one of the collinear variables our model would decline in performance, so we decided to try our multiplying the corresponding collinear variables to gether and BINGO! We got and R^2 of .3247 using the following model:

```
TARGET_WINS ~ BATTING_3B + BATTING_HR + BATTING_BB*PITCHING_BB +
    BATTING_SO*PITCHING_SO + BASERUN_SB + BASERUN_CS + BATTING_H*log(PITCHING_H) +
    log(FIELDING_E) + FIELDING_DP
```

3.3.4.2 R^2 0.3247108

3.3.5 Predictions

We ran predictions on our final model and plotted the distribution next to the distribution from our target in the training data set to compare...

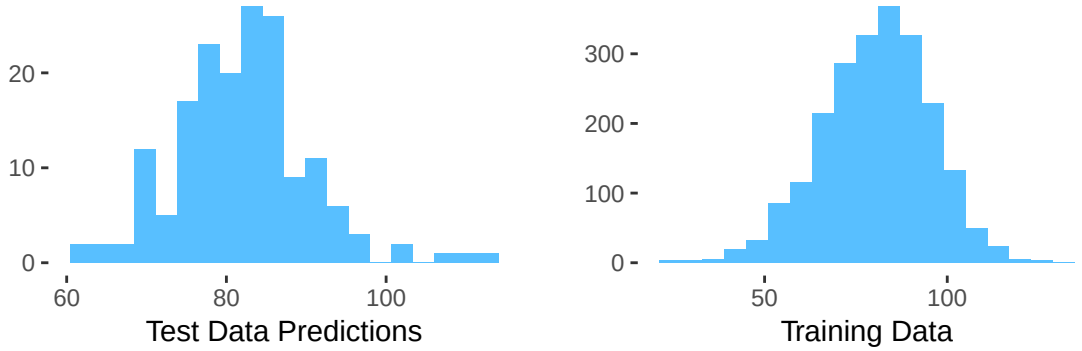


Figure 12: Predictions vs. training data

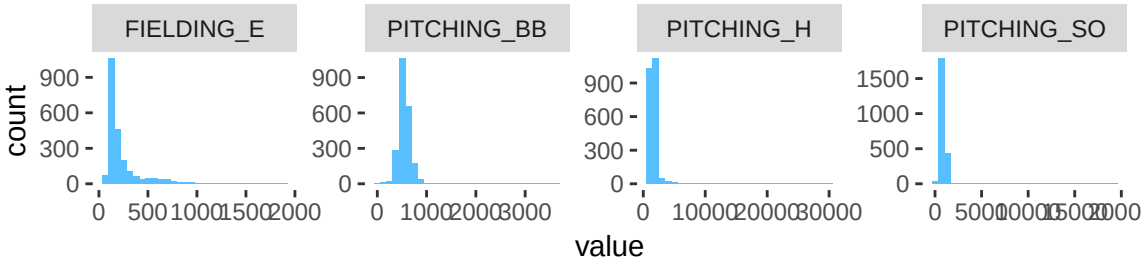


Figure 13: Histograms of variables showing pronounced rightward-skew

3.4 MODEL 3

We sought to explore whether there was a relationship between wins and the difference of specific offensive and defensive team capabilities - hits, homeruns, balls, and strike-outs. Incorporating variables that reflect those differences (i.e. subtracting batting hits from pitching hits, and so on), however, did not improve the explantory power of the model beyond using the original variables.

Given these variables did not yield improvements, in their place we explored a third model. As the histograms below highlight, a number of the independent variables - pitching hits, pitching homeruns, pitching strikeouts - demonstrate pronounced rightward-skew.

We corrected for that skew by transforming those three variables using natural logarithms. When we tested those log transformations in a model where they replaced the untransformed, original variables combined with all other variables, we found that neither the originals nor the log transformations for pitching homeruns and pitching strikeouts met the threshold of significance (a p-value below the α level of .05). Based on high p-values, over a series of backward steps we removed pitching homeruns, pitching strikeouts, and baserun caught stealing, yielding the following model:

[Jeremy: team, should we write LaTeX formulas for each model or just cable model coefficients?]

$y = $

Based on this model's F-statistic and p-value, we can reject the null hypothesis that coefficients with values of zero would fit the data better. Per the adjusted r^2 value, this model explains approximately 29.56% of the variance in wins. However, in doing so it treats the batting hits and batting second base runs as drags on wins (with negative coefficients), and pitching hits as buoying wins - which is counterunituitive. While the other coefficients make more intuitive sense, these signs call into question how effectively we can use this model to understand the relationships between the independent variables and wins.

Table 6: Log Transform Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-405.0139425	36.9977658	-10.946984	0.0000000
BATTING_H	-0.0134654	0.0060337	-2.231679	0.0257358
BATTING_2B	-0.0150055	0.0091463	-1.640604	0.1010214
BATTING_3B	0.1405383	0.0176556	7.959979	0.0000000
BATTING_HR	0.0723000	0.0097039	7.450614	0.0000000
BATTING_BB	0.1246742	0.0126620	9.846362	0.0000000
BATTING_SO	-0.0065541	0.0023297	-2.813291	0.0049469
BASERUN_SB	0.0438045	0.0047630	9.196774	0.0000000
log(PITCHING_H)	68.6642334	5.7601412	11.920582	0.0000000
PITCHING_BB	-0.0951797	0.0106315	-8.952603	0.0000000
FIELDING_E	-0.0473986	0.0035504	-13.350085	0.0000000
FIELDING_DP	-0.0885260	0.0129387	-6.841948	0.0000000

---

4 SELECT MODELS

4.1 Instructions:

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

4.2 Comparison of models

[Jeremy: added in a chart for model 3]

5 Multi-collinearity

[Jeremy: We can place the correlelogram here or revisit it]

An examination of the partial correlation coefficients between all the independent variables shows that there the expected relationships between the offense and defense variables - i.e. batting and pitching homeruns - are there, meet p-value thresholds, and are strong. This finding suggests keeping only one of each pair of variables in a given model; yet, when we tried this, less of the variance in wins was explained by the model.

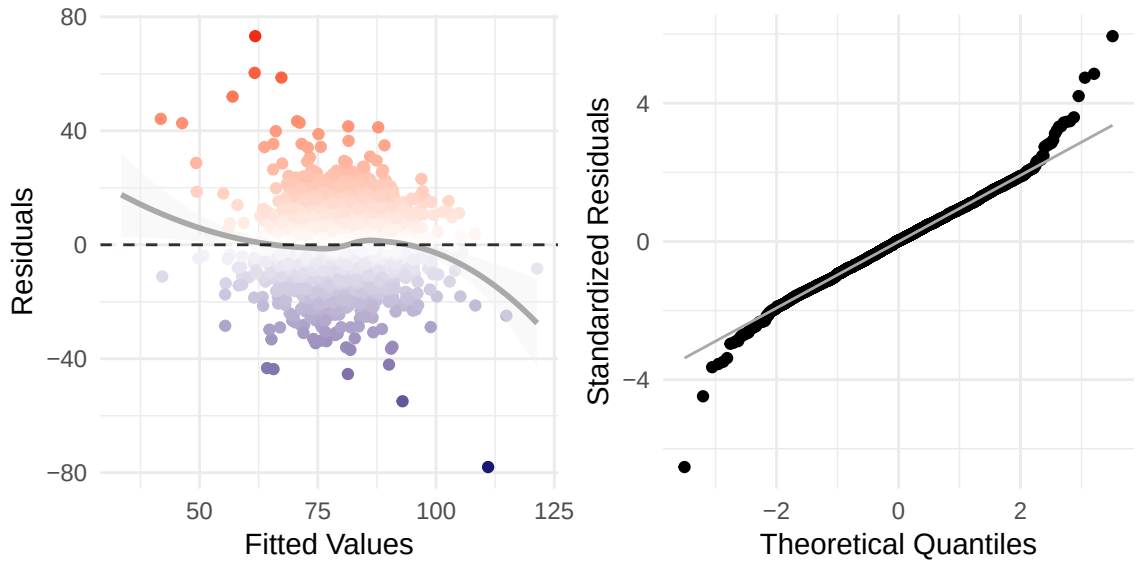


Figure 14: Model 3: Residual Plot and Q-Q Plot

	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN
BATTING_2B	1.0000000	0.2129669	0.0609690	0.1206106	-0.0764713	0.107
BATTING_3B	0.2129669	1.0000000	-0.0730874	0.0131362	-0.1506443	0.171
BATTING_HR	0.0609690	-0.0730874	1.0000000	0.5747337	0.1124715	0.015
BATTING_BB	0.1206106	0.0131362	0.5747337	1.0000000	0.5064157	-0.160
BATTING_SO	-0.0764713	-0.1506443	0.1124715	0.5064157	1.0000000	0.337
BASERUN_SB	0.1073938	0.1713437	0.0156533	-0.1602112	0.3375119	1.000
BASERUN_CS	-0.1126072	0.0577696	-0.0273953	-0.0008126	-0.1012046	0.617
PITCHING_H	0.3500087	-0.1142022	0.1033127	-0.3937551	-0.0524595	-0.079
PITCHING_HR	-0.0287742	0.0540702	0.9817947	-0.5681054	-0.0702943	-0.000
PITCHING_BB	-0.1060008	-0.0071098	-0.5899436	0.9778943	-0.5319800	0.225
PITCHING_SO	0.0983961	0.0219166	-0.0598031	-0.5935317	0.9486979	-0.308
FIELDING_E	-0.2553727	0.1729403	0.2509843	-0.0913065	-0.2047899	0.220
FIELDING_DP	0.0551834	0.0042580	-0.0212282	0.0455742	-0.0470879	-0.013

	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN
BATTING_2B	0.0000000	0.0000000	0.0040650	0.0000000	0.0003115	0.000
BATTING_3B	0.0000000	0.0000000	0.0005699	0.5362626	0.0000000	0.000
BATTING_HR	0.0040650	0.0005699	0.0000000	0.0000000	0.0000001	0.461
BATTING_BB	0.0000000	0.5362626	0.0000000	0.0000000	0.0000000	0.000
BATTING_SO	0.0003115	0.0000000	0.0000001	0.0000000	0.0000000	0.000
BASERUN_SB	0.0000004	0.0000000	0.4611241	0.0000000	0.0000000	0.000
BASERUN_CS	0.0000001	0.0064878	0.1970481	0.9694840	0.0000018	0.000
PITCHING_H	0.0000000	0.0000001	0.0000011	0.0000000	0.0134555	0.000
PITCHING_HR	0.1754304	0.0108506	0.0000000	0.0000000	0.0009212	0.978
PITCHING_BB	0.0000006	0.7378265	0.0000000	0.0000000	0.0000000	0.000
PITCHING_SO	0.0000034	0.3020954	0.0048318	0.0000000	0.0000000	0.000
FIELDING_E	0.0000000	0.0000000	0.0000000	0.0000165	0.0000000	0.000
FIELDING_DP	0.0093223	0.8411154	0.3175375	0.0318142	0.0265469	0.533

	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN
BATTING_2B	0.000000	10.2629918	2.8760766	5.7207133	-3.611228	5.086
BATTING_3B	10.262992	0.0000000	-3.4505512	0.6185709	-7.174974	8.188
BATTING_HR	2.876077	-3.4505512	0.0000000	33.0685744	5.329542	0.737
BATTING_BB	5.720713	0.6185709	33.0685744	0.0000000	27.652657	-7.642
BATTING_SO	-3.611228	-7.1749738	5.3295425	27.6526571	0.000000	16.882
BASERUN_SB	5.086057	8.1888267	0.7371254	-7.6422669	16.882395	0.000
BASERUN_CS	-5.336053	2.7246343	-1.2903952	-0.0382598	-4.789815	36.969
PITCHING_H	17.592987	-5.4126239	4.8906535	-20.1693441	-2.473462	-3.752
PITCHING_HR	-1.355393	2.5496259	243.3742845	-32.5038836	-3.318017	-0.027
PITCHING_BB	-5.019328	-0.3347749	-34.4017617	220.2020200	-29.581449	10.907
PITCHING_SO	4.655576	1.0321907	-2.8208809	-34.7242694	141.276857	-15.288
FIELDING_E	-12.436593	8.2674688	12.2083774	-4.3172042	-9.851330	10.665
FIELDING_DP	2.602278	0.2004905	-0.9997564	2.1480957	-2.219598	-0.622